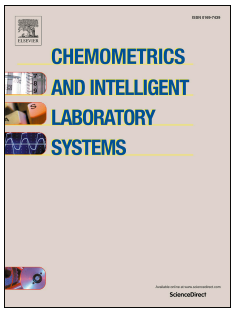
Accepted Manuscript



A model-based data mining approach for determining the domain of validity of approximated models

Marco Quaglio, Eric S. Fraga, Enhong Cao, Asterios Gavriilidis, Federico Galvanin

PII: S0169-7439(17)30448-3

DOI: 10.1016/j.chemolab.2017.11.010

Reference: CHEMOM 3543

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 5 July 2017

Revised Date: 13 October 2017

Accepted Date: 11 November 2017

Please cite this article as: M. Quaglio, E.S. Fraga, E. Cao, A. Gavriilidis, F. Galvanin, A model-based data mining approach for determining the domain of validity of approximated models, *Chemometrics and Intelligent Laboratory Systems* (2017), doi: 10.1016/j.chemolab.2017.11.010.

A model-based data mining approach for determining the domain of validity of approximated models

Marco Quaglio[a], Eric S. Fraga[a], Enhong Cao[a], Asterios Gavriilidis[a], Federico Galvanin[a,*]

[a] *Department of Chemical Engineering, University College London (UCL), Torrington Place, WC1E 7JE London, United Kingdom*

Abstract

Parametric models derived from simplifying modelling assumptions give an approximated description of the physical system under study. The value of an approximated model depends on the consciousness of its descriptive limits and on the precise estimation of its parameters. In this manuscript, a framework for identifying the model domain of validity for the simplifying model hypotheses is presented. A model-based data mining method for parameter estimation is proposed as central block to classify the observed experimental conditions as compatible or incompatible with the approximated model. A nonlinear support vector classifier is then trained on the classified (observed) experimental conditions to identify a decision function for quantifying the expected model reliability in unexplored regions of the experimental design space. The proposed approach is employed for determining the domain of reliability for a simplified kinetic model of methanol oxidation on silver catalyst.

Keywords: model identification, maximum likelihood, data mining, machine learning, model diagnosis

*Corresponding author
Email address: f.galvanin@ucl.ac.uk (Federico Galvanin)

1. Introduction

The exhaustive description of most biochemical and physicochemical processes requires the development of complex models involving systems of differential and algebraic equations. The identification of detailed model structures is frequently hindered by limitations in the experimental setup (e.g. impossibility of measuring some physical quantities or separating mechanisms with overlapping effects), and/or prohibitive experimental cost. In these situations, "lumped" models derived from simplifying hypotheses are normally proposed, fitted to the experimental data and tested with statistically appropriate methods, e.g. a χ^2-test. A failed χ^2-test is interpreted as an incorrect or incomplete set of modelling hypotheses and the modelling activity may proceed in two different ways:

1. new formulations of the model, are proposed, tested and compared adopting techniques of model building available in the literature [1];
2. the incorrect model structure may be maintained accepting its limited capabilities of describing the physical reality under analysis.

The present manuscript focuses on the second approach to phenomenological modelling and on how to improve the predictive capabilities of approximated model structures.

The identification of an approximated model, once a suitable model structure is selected, requires:

- the precise estimation of the model parameters through the fitting of experimental data carrying valuable information;
- the identification of the range of experimental conditions for which the model can provide reliable predictions, i.e., the domain of validity of the model hypotheses.

Optimal experimental conditions for the estimation of the model parameters can be identified employing model-based design of experiments (MBDoE) techniques for parameter precision [1, 2]. MBDoE methods intrinsically assume that

the model structure is reliable, i.e., the model is assumed to provide a good fitting and good predictions all across the experimental design space. However, this assumption may not be acceptable in the presence of an approximated model structure. In these situations, the research of the optimal experimental conditions should be bounded within the domain of validity of the simplifying modelling assumptions.

In this work, a framework for the identification of approximated models is proposed. The investigated experimental conditions are labelled as compatible or incompatible with the modelling hypotheses at the stage of parameter estimation by employing a model-based data mining tool derived from the maximum likelihood method [3]. The labelling is then used to train a supervised machine learning algorithm based on support vector theory [4, 5] to map unexplored experimental conditions in terms of satisfactory or unacceptable expected model performance. The generated map can then be employed for preventing the use of false optimal process points located in regions of low model reliability or for supporting the design of new trials to enhance parameter precision.

2. Methodology

Assume that a model derived from simplifying hypotheses is proposed for interpreting a certain physical system.

$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t, \boldsymbol{\theta}) \tag{1}$$

In Eq. (1) $\hat{\mathbf{y}}$ represents an N_m-dimensional array of measurable model outputs, $\mathbf{x}$ is an N_x-dimensional vector of state variables, $\mathbf{u} \in U$ is an N_u-dimensional vector of control variables and t is time. $\boldsymbol{\theta} \in \Theta$ represents an array of N_θ model parameters that require estimation. Assume that a number of experiments N_{exp} are performed at experimental conditions $\mathbf{u}_j$ with $j = 1, ..., N_{exp}$ obtaining a preliminary set of data $\Psi = \{y_{ij} | i = 1, ..., N_m; j = 1, ..., N_{exp}\}$ and that measurements are characterised by uncorrelated Gaussian noise with known standard deviations σ_{ij}. A standard approach for estimating model parameters

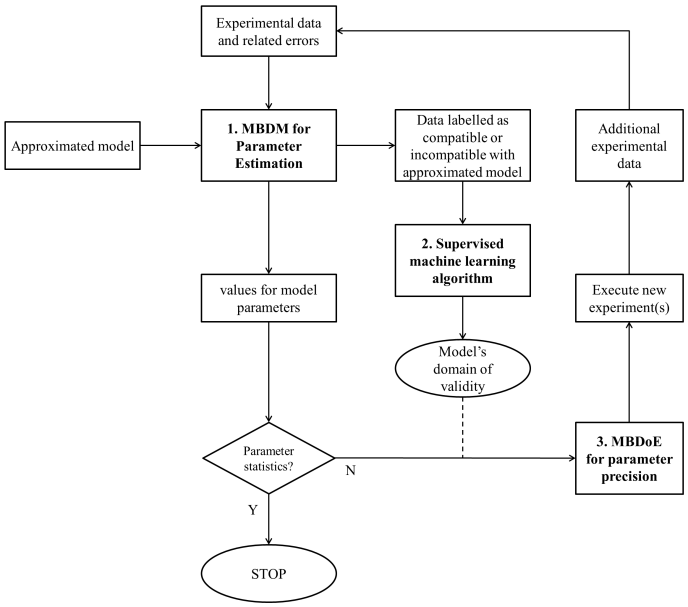


Figure 1: Proposed framework for model identification. Boldface blocks represent fundamental steps in the proposed methodology. The procedure starts from the availability of a preliminary set of experimental data and an approximated model structure to describe the phenomenon. Non-measurable model parameters are estimated fitting the available dataset through Model-Based Data Mining (MBDM) methods for parameter estimation (1). MBDM computes an instance for the model parameters and classifies the observed experimental conditions as compatible or incompatible with the proposed model. The labelling computed by MBDM is then processed by a supervised machine learning algorithm to extend the classification to unexplored regions of the design space (2). The training of the learning machine leads to the determination of a model's domain of validity for the modelling assumptions. A check on the statistical quality of the parameters computed by MBDM is then performed and, in case of statistically unsatisfactory estimates, additional experiments are designed through Model-Based Design of Experiment (MBDoE) methods for parameter precision (3). The research of optimal experimental conditions to investigate is bounded to the model's domain of validity to prevent the collection of model-incompatible experimental data.

from experimental data is the least square method. However, least squares approaches do not account properly for measurement noise in the parameter estimation. A more sophisticated method that demonstrated to provide good estimates in a broad range of situations is the maximum likelihood estimator [3]. The method derives from the assumption that it does exist a value of the parameters, namely the maximum likelihood (ML) estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, which maximises the likelihood of observing the experimental data, given the model parametrisation. The computation of the ML estimate is performed through the maximisation of the likelihood function L or, indifferently, its natural logarithm $\Phi_{\mathrm{ML}} = \ln L$ [3].

$$\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m} -\ln(2\pi\sigma_{ij}^2) - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 \tag{2}$$

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi) \tag{3}$$

A discrepancy between the distribution of model residuals $\hat{y}_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - y_{ij}$ and the distribution of the measurement errors is interpreted as a consequence of incorrect model specification, and it is normally detected through statistical tests that assess the goodness of fit (e.g. a χ^2-test).

Conventional estimators (e.g. least squares or ML) do not take into account the structural uncertainty on the model equations. If the model structure is approximated, one shall not expect the model to give good predictions throughout the whole experimental design space and for all its measurable output variables. As a direct consequence, not all the collected data may be significant for the estimation of the model parameters. In this work, a framework for the identification of approximated models is proposed to address the multi-objective task of both parameter estimation and the determination of the model domain of reliability. The method follows from the assumption that any model structure is capable of fitting accurately experimental data as long as the fitted domain is not excessively vast (as an example, any continuous nonlinear function is locally well approximated by a linear model). The model identification framework, represented in Figure 1, involves three fundamental steps:

1. *A Model-based data mining step for parameter estimation.* At this stage, the experimental data for which the candidate model is unable to realise low residuals are identified and excluded from the parameter estimation problem employing model-based data mining (MBDM) methodologies. MBDM produces two outputs: *i*) it generates an instance of parameters for the candidate model structure and *ii*) it labels the data as compatible or incompatible with the modelling hypotheses.
2. *A supervised machine learning training step.* The labelling of the data generated at step 1 by MBDM is processed by a supervised machine learning algorithm, e.g. a Support Vector Machine (SVM) [5, 6], in order to map the experimental design space in terms of good and bad expected model predictive capabilities.
3. *A MBDoE step for parameter precision.* If some parameter estimates are found to be statistically unsatisfactory, new data have to be collected and included in the parameter estimation problem. Model-based design of experiments (MBDoE) methods for parameter precision can be employed at this stage to identify highly informative experimental conditions within the range of expected model reliability identified at step 2.

In the following sections, model-based data mining methods derived from the maximum likelihood approach are presented to address task 1. The underlying mathematics of SVM technology is then presented with the aim of addressing task 2. It is not in the aims of this manuscript to present and detail MBDoE methods for parameter precision, for which an extensive literature is available [7–10].

2.1. Model-Based Data Mining for Parameter Estimation

The approach illustrated here is proposed with the aim of addressing the problem of parameter estimation through the automated selection of model-compatible experimental data. Model-compatible data represent a subset $\Psi' \subseteq \Psi$ of the whole available dataset such that the fitting of Ψ' leads to a distribution of model residuals that cannot be distinguished from a distribution of

measurement errors. The necessity of making an assumption on the distribution of the measurement noise justifies the employment of a ML approach as a starting point for the following derivations.

It is assumed that measurements are characterised by Gaussian noise with known standard deviations σ_{ij} and that σ_{ij} do not depend on $\boldsymbol{\theta}$. Thus, the location of the ML estimate in Θ, which depends on the gradient of (2), is not influenced by the magnitude of elements $-\ln(2\pi\sigma_{ij}^2)$. Elements $-\ln(2\pi\sigma_{ij}^2)$ in Eq. (2) may be therefore substituted with arbitrary constants c_{ij} without affecting the result of the parameter estimation problem (3). If $c_{ij} = z_{\frac{\alpha}{2}}^2$ ($\forall$ $i = 1, ..., N_m$ and $j = 1, ..., N_{exp}$), where $z_{\frac{\alpha}{2}}$ is the two-tailed z-value with significance α derived from a normal standard distribution, the log-likelihood function becomes:

$$\Phi'_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m} z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 \tag{4}$$

In Eq. (4) the ij-th element brings a positive contribution to the function only if:

$$|\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}| < z_{\frac{\alpha}{2}}\sigma_{ij} \tag{5}$$

One may decide to exclude from the objective function the data that do not satisfy condition (5). For this purpose, an N_mxN_{exp} matrix $\boldsymbol{\Lambda}$ of binary variables $\lambda_{ij} \in \{0, 1\}$ is defined and the log-likelihood in Eq. (4) is modified as follows:

$$\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m} \lambda_{ij}\left[z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2\right] \tag{6}$$

$$\text{s.t.} \quad \lambda_{ij} = \begin{cases} 1 & \text{if} \quad z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 \geqslant 0 \\ 0 & \text{if} \quad z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 < 0 \end{cases} \quad \forall \quad i, j \tag{7}$$

In Eq. (6), binary variables λ_{ij} act like switchers excluding the data whose contribution to $\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi)$ is negative. It is important to notice that conditions (7) state the dependence of the binary variables on the values of the parameters $\boldsymbol{\theta}$. The parameter estimation problem is reformulated as:

$$\hat{\boldsymbol{\theta}}_{\mathrm{DM}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi) \tag{8}$$

Solution $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ is the result of the fitting of a potentially reduced set of measurements $\Psi' \subseteq \Psi$, i.e. only the measurements for which the model can provide low residuals that satisfy (5).

$$\Psi' = \{y_{ij} | \lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 1\} \tag{9}$$

If $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ maximises Eq. (6) then it also identifies the global optimum of $\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi')$, i.e. the log-likelihood function involving the reduced set of measurements, indeed:

$$\begin{aligned}\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi') = \Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi') + \Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi) - \Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi) \\ + \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}})z^2_{\frac{\alpha}{2}} - \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}})z^2_{\frac{\alpha}{2}}\end{aligned} \tag{10}$$

$$\begin{aligned}\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi') = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\boldsymbol{\theta})\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \\ + \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\big(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\big)\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \\ + \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}})\Big[-\ln(2\pi\sigma^2_{ij}) - z^2_{\frac{\alpha}{2}}\Big]\end{aligned} \tag{11}$$

The first sum in Eq. (11) is $\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi)$ and it reaches its global maximum in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$. The second sum is always non-positive (because of conditions (7)) and it is null for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{DM}}$. The third sum is constant and it does not depend on parameters. Thus, it is concluded that $\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi') \leq \Phi_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}|\Psi')\ \forall\ \boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}_{\mathrm{DM}}$. A more detailed demonstration of this conjecture is given in Appendix A.

The presented approach can be employed to prompt the exclusion of entire experiments rather than single data. For this purpose and for reasons that will become clear in the next section, binary variables $\beta_j \in \{1, -1\}$ with $j = 1, ..., N_{exp}$ are defined and (4) is modified as follows.

$$\Phi_{\mathrm{DM},exp}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\Big(\frac{1+\beta_j}{2}\Big)\sum_{i=1}^{N_m}z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2 \tag{12}$$

$$\text{s.t.} \quad \beta_j = \begin{cases} +1 & \text{if} \quad \sum_{i=1}^{N_m} z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 \geqslant 0 \\ -1 & \text{if} \quad \sum_{i=1}^{N_m} z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\right)^2 < 0 \end{cases} \quad \forall \quad j \tag{13}$$

$$\hat{\boldsymbol{\theta}}_{\mathrm{DM},exp} = \arg\max_{\boldsymbol{\theta}\in\Theta} \Phi_{\mathrm{DM},exp}(\boldsymbol{\theta}|\Psi) \tag{14}$$

Analogously to (8), Eq. (12) can then be optimised with respect to the model parameters. Given a reasonable choice for the significance α for the $z_{\frac{\alpha}{2}}$-value, the solution of (14) leads to the automated exclusion from the parameter estimation problem of the experiments that are incompatible with the modelling assumptions, i.e.: *1*) experiments performed outside the domain of model reliability and *2*) experiments in which measurements are affected by excessive error. Notice that MBDM does not distinguish between these two categories. In fact MBDM only classifies the experiments based on the associated fitting realised by the candidate model. A possible practical way to provide a more accurate classification of the data in the two aforementioned categories is to repeat the experiments. If the incompatibility persists after the repetition, the experiment shall be classified in the first category, i.e. the trial was performed outside the domain of model reliability. If the repeated experiment is instead found to be compatible, the incompatibility detected before the repetition shall be interpreted as a consequence of an excessive measurement error.

2.2. Identification of a region of model reliability

The solution of the optimisation problem (14), leads to the construction of a function $\psi : \{\mathbf{u}_j | j = 1, ..., N_{exp}\} \rightarrow \hat{\beta}_j \in \{1, -1\}$ (where $\hat{\beta}_j = \beta_j(\hat{\boldsymbol{\theta}}_{\mathrm{DM},exp})$), which classifies the explored experimental conditions $\mathbf{u}_j$, with $j = 1, ..., N_{exp}$, either as compatible or incompatible with the candidate model. It is now of interest to identify a decision function $I(\mathbf{u})$, based on the training set $\{(\mathbf{u}_j, \hat{\beta}_j) | j = 1, ..., N_{exp}\}$, whose sign can be used to classify the performance of the model in unexplored experimental conditions. A decision function is required for quantifying: *i*) the reliability on the model predictions across the model input space

U; *ii*) the expected model fitting across the experimental design space for supporting the design of new trials to enhance parameter precision. The problem may be recast in terms of identifying a hyperplane in the input space U that classifies the training set with the minimum misclassification error. Let the generic hyperplane in the input space U be:

$$\mathbf{w}^T\mathbf{u} + b = 0 \tag{15}$$

where $\mathbf{w}$ is an N_u-dimensional array of coefficients and the scalar quantity b represents the offset of the hyperplane. From SVM theory, it is known that the optimal hyperplane can be identified solving the following convex optimisation problem [4]:

$$\begin{aligned} &\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\Big(\sum_{j=1}^{N_{exp}} \xi_j\Big)^2 \\ &\text{s.t.} \quad \hat{\beta}_j(\mathbf{w}^T\mathbf{u}_j + b) \geqslant 1 - \xi_j, \\ &\qquad \xi_j \geqslant 0, \quad j = 1, ..., N_{exp} \end{aligned} \tag{16}$$

where C is a regularisation constant and vector $\boldsymbol{\xi} = \{\xi_j | j = 1, ..., N_{exp}\}$ accounts for misclassification errors. However, since both the physical system and the candidate model equations may be highly nonlinear one shall not expect a hyperplane in the input space to provide a good classification. In fact, in general, the domain of model reliability may be non-convex, non-compact and finite. In order to represent a potentially very complicated geometry in the input space, the so called *kernel trick* [5] is employed. The basic idea is to map the training set through a nonlinear transformation $\boldsymbol{\varphi} : U \to Z$ into a feature space where the separation through a hyperplane becomes more significant. It is then possible to demonstrate that the decision function $I(\mathbf{u})$ in the input space has the form:

$$I(\mathbf{u}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{u}) + b = \sum_{j=1}^{N_{exp}} a_j\hat{\beta}_j K(\mathbf{u}, \mathbf{u}_j) + b \tag{17}$$

where a_j with $j = 1, ..., N_{exp}$ are the Lagrange multipliers derived from the maximisation of the unconstrained Lagrange form of problem (16) and $K(\mathbf{u}, \mathbf{u}_j) =$

$\boldsymbol{\varphi}(\mathbf{u})^T\boldsymbol{\varphi}(\mathbf{u}_j)$ represents a specific kernel function. Notice that in order to compute $I(\mathbf{u})$ it is not necessary to know the form of $\boldsymbol{\varphi}$, but only its associated kernel, which must be chosen *a priori*. A popular choice for K is the Gaussian radial basis function:

$$K(\mathbf{u}, \mathbf{u}_j) = e^{-\frac{(\mathbf{u}-\mathbf{u}_j)^T(\mathbf{u}-\mathbf{u}_j)}{2\gamma^2}} \tag{18}$$

where γ is a tuning parameter that can be interpreted as decay length of the radial function and determines the degree of *similarity* between two different sets of experimental conditions. Since SVM are sensitive to the scale of the input space, it is recommended to normalise the experimental conditions before the application of the learning machine. Notice that if a radial basis function is selected as kernel, two degrees of freedom are present due to the presence of the regularisation constant C (which trades off smoothness of the decision surface and misclassification) and the tuning parameter γ. The hyperparameters C and γ may be chosen *a priori* or, in the presence of sufficiently large data sets, an optimal hyperparameter set may be identified through cross-validation [11].

2.3. Model-Based Design of Experiments for Parameter Precision

Conventional MBDoE methods for parameter precision do not take into account the expected accuracy on the model predictions, i.e., an unconstrained MBDoE problem may lead to the design of sampling points outside the domain of validity of the model. The reliability function (17) can be fruitfully employed in a model-based experimental design framework to bound the research of optimal informative experimental conditions in regions of U where $I(\mathbf{u}) > 0$ (i.e. the regions of the design space where the model is expected to provide a good fitting). As previously mentioned, the present manuscript is focused on the description and application of steps 1 and 2 of the proposed framework for model identification (see Figure 1). A detailed description of MBDoE methods for parameter precision can be found in the literature [7–10].

3. Case study

The presented methodology is now employed to identify a domain of expected model reliability for an approximated kinetic model proposed to describe methanol oxidation on silver catalyst in continuous flow microreactors [12, 13]. A short presentation of the experimental setup and the available data set is followed by a description of the candidate model. Eventually, the assumptions made for the application of MBDM and SVM are presented.

3.1. Experimental setup and data set

Microfluidic devices are promising means for gathering information on chemical kinetics. Due to their small dimensions, reactions can be conducted in the absence of heat and mass transfer resistances [14, 15]. A data set Ψ consisting of 13 steady-state kinetic experiments was collected on a silicon-glass microreactor. A schematic diagram of the device is given in Figure 2. The reactor chip was constructed from a silicon wafer through photolithography and deep reactive ion etching. A thin layer of silver was sputtered on the bottom of the microchannel obtaining a catalyst film 78.1 mm in length. Mass flow controllers were used to inject the gaseous mixture consisting of methanol, oxygen, water and helium (added as inert diluent). A detailed description of the setup is available in the literature [16]. The explorable design space U in the setup consists of five independent input variables: temperature T of the microreactor; flowrate F of the gaseous mixture at the inlet; molar fractions of methanol, oxygen and water in the inlet mixture, i.e., $y^{IN}_{\mathrm{CH_3OH}}$, $y^{IN}_{\mathrm{O_2}}$ and $y^{IN}_{\mathrm{H_2O}}$ respectively. The experiments were conducted varying one factor at time to assess the effect on the outlet composition. A summary of the investigated experimental conditions is given in Table 1. The main products of the reaction in the investigated range of conditions are: formaldehyde, carbon dioxide, hydrogen and water. The composition of the mixture at the outlet was analysed through gas chromatography.

Table 1: Experimental conditions investigated in the available experiments. The volumetric flowrate F is referred to standard conditions. Helium, used as inert carrier, represents the remaining molar fraction at the inlet.

N° exp.	T [K]	F* [ml min^{-1}]	$y^{IN}_{\mathrm{CH_3OH}}$	$y^{IN}_{\mathrm{O_2}}$	$y^{IN}_{\mathrm{H_2O}}$
1-3	783	29.1-73.1	0.0996	0.0414	0.0754
4-7	733-826	50.9	0.1468	0.0975	0.2293
8-10	765-826	93.9	0.1469	0.0980	0.2296
11-13	800-900	54.5	0.2590	0.1064	0.2122

* at temperature $T = 273.15$ K; pressure $P = 101325$ Pa.

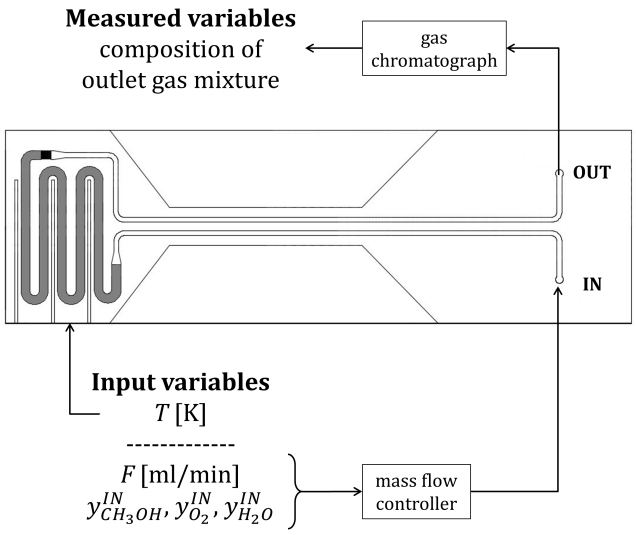


Figure 2: Schematic representation of the microreactor chip and setup. The grey-coloured area in the microchannel represents the sputtered silver catalyst film.

3.2. Modelling assumptions

The section of the microchannel occupied by the silver catalyst film is modelled as an ideal plug-flow reactor. Isothermal conditions are assumed to be realised along the whole length of the channel (i.e. the energy balance is omit-

ted), and diffusion phenomena are completely neglected. The generic form of the mass balance is given in (19), where N_C and N_R represent the number of components and the number of reactions respectively, C_i is the species concentration expressed in $\mathrm{mol\,m^{-3}}$, z represents the axial coordinate of the channel in m, v is the flow velocity along z expressed in $\mathrm{m\,s^{-1}}$, ν_{ij} is the stoichiometric coefficient of the i-th component in the j-th reaction and r_j is the rate associated to the j-th reaction, expressed in $\mathrm{mol\,m^{-3}\,s^{-1}}$.

Mass balances

$$v\frac{dC_i}{dz} = \sum_{j=1}^{N_R} \nu_{ij} r_j \quad \forall \quad i = 1, ..., N_C \tag{19}$$

A simplified kinetic model derived from the one proposed by Andreasen *et al.* [12], is adopted to model the reaction. Andreasen's model assumes the presence of two limiting steps: 1) a step of oxidative dehydrogenation of methanol to formaldehyde and 2) a step of complete oxidation of formaldehyde to carbon dioxide.

Stoichiometry and Kinetics

$$\begin{aligned}
&\text{Reaction 1:} && \mathrm{CH_3OH} + \tfrac{1}{4}\mathrm{O_2} \leftrightarrow \mathrm{CH_2O} + \tfrac{1}{2}\mathrm{H_2} + \tfrac{1}{2}\mathrm{H_2O} \\
& && r_1 = A_1 e^{-\frac{E_{a1}}{RT}} \frac{C_{\mathrm{CH_3OH}} C_{\mathrm{O_2}}^{0.25}}{C_{\mathrm{H_2O}}^{0.5}} \\
&\text{Reaction 2:} && \mathrm{CH_2O} + \tfrac{1}{2}\mathrm{O_2} \leftrightarrow \mathrm{H_2} + \mathrm{CO_2} \\
& && r_2 = A_2 e^{-\frac{E_{a2}}{RT}} \frac{C_{\mathrm{CH_2O}} C_{\mathrm{O_2}}^{0.5}}{C_{\mathrm{H_2}}^{0.5}} \\
&\text{Reaction 3:} && \mathrm{H_2} + \tfrac{1}{2}\mathrm{O_2} \rightarrow \mathrm{H_20} \\
& && r_3 = A_3 e^{-\frac{E_{a3}}{RT}} C_{\mathrm{H_2}} C_{\mathrm{O_2}}^{0.5}
\end{aligned} \tag{20}$$

A reaction of hydrogen oxidation is also included in the kinetic model. Hydrogen oxidation is known to occur only at higher temperature [17], and it was chosen to include it in the model primarily for describing the low amounts of

hydrogen detected at the outlet. A total of $N_C = 6$ species are considered in the approximated kinetics, i.e., methanol, oxygen, water, formaldehyde, hydrogen and carbon dioxide. The rates of the three reactions are given in (20), where R is the ideal gas constant, A_j and E_{aj} (with $j = 1, ..., 3$) represent pre-exponential factors and activation energies of the Arrhenius type rate constants. An instance for the kinetic parameters was available from previous kinetic investigations, conducted on a different setup [18]. The values are reported in Table 2. Since the reactivity of the catalyst film is highly influenced by its fabrication history, one shall not expect the parameter instance given in Table 2 to be representative for the catalyst employed in this case study. The different kinetic behaviour between different silver catalyst types is assumed to derive from a different density of active sites on the film surface. Following this assumption, only the pre-exponential factors of the catalytic reactions shall be tuned on the available data set Ψ. The catalyst promotes the partial oxidation of methanol and the oxidation of formaldehyde, i.e. Reaction 1 and Reaction 2. Evidence of catalytic influence of silver on hydrogen oxidation, i.e. Reaction 3, is reported in the literature [19]. However, for the purposes of this work, Reaction 3 is assumed to be independent from the catalyst behaviour, i.e., a different density of active sites on the catalyst surface does not influence the kinetic rate of hydrogen oxidation. Thus, in this case study, A_3, E_{a1}, E_{a2} and E_{a3} are fixed to the values given in Table 2 while A_1 and A_2 are treated as the parameters requiring estimation, i.e. $\boldsymbol{\theta} = [A_1, A_2]$.

3.3. Methods

Since the model presented in Section 3.2 was derived by a number of simplifying hypotheses, its identification requires both the quantification of the unknown parameters $\boldsymbol{\theta} = [A_1, A_2]$ through data fitting, and the identification of a region of reliability in the input space associated to the estimated parameters. The task is fulfilled through the following steps:

1. The set of parameters $\boldsymbol{\theta}$ is estimated employing MBDM (14) fitting the molar fractions of methanol, oxygen, water, formaldehyde, hydrogen and

Table 2: Instance for the kinetic parameters obtained from previous kinetic studies [18].

Parameter	Unit	Value
A_1	$[(\text{mol m}^{-3})^{0.25}\text{s}^{-1}]$	$5.33 \cdot 10^{11}$
A_2	$[\text{s}^{-1}]$	$1.03 \cdot 10^{7}$
A_3	$[(\text{mol m}^{-3})^{-0.5}\text{s}^{-1}]$	$1.07 \cdot 10^{4}$
E_{a1}	$[\text{J mol}^{-1}]$	$1.42 \cdot 10^{5}$
E_{a2}	$[\text{J mol}^{-1}]$	$9.02 \cdot 10^{4}$
E_{a3}	$[\text{J mol}^{-1}]$	$1.83 \cdot 10^{4}$

carbon dioxide detected at the outlet in the 13 experiments, i.e, $N_m = 6$ and $N_{exp} = 13$. The measured molar fractions are assumed to be affected by Gaussian noise with $\sigma_{ij} = 3 \cdot 10^{-3}\ \forall\, i, j$. The tuning constant is set at $z_{\frac{\alpha}{2}} = 3$ (which corresponds to a significance $\alpha = 0.997$).

2. Binary variables $\hat{\beta}_j$ (with $j = 1, ..., N_{exp}$) obtained by MBDM at step 1 are used to generate the training set. The explored range of experimental conditions is normalised to the unit hypercube in the input space U for the application of the SVM.
3. A SVM is employed to identify a reliability decision function $I(\mathbf{u})$ in the experimental design space. Two cases are considered:
   - *Case 1:* the model is assumed to be weak at describing certain ranges of temperature and inlet fraction of methanol while it is assumed to be reliable on the other experimental conditions. The SVM machine is therefore trained assuming a bi-dimensional input space $U = (T, y^{IN}_{\text{CH}_3\text{OH}})$;
   - *Case 2:* the model is considered weak in representing the system in broad ranges of temperature and inlet fraction of water, but reliable on other experimental conditions. The SVM machine is then trained on the reduced input space $U = (T, y^{IN}_{\text{H}_2\text{O}})$.

Since in the present case study the training set involves only 13 points, it is

chosen to set the hyperparameters of the learning machine *a priori* instead of determining them through cross-validation. In both cases, a Gaussian kernel is employed (18) with $\gamma = 0.2$; being the experimental conditions in the training set normalised, this corresponds to having a characteristic decay length equal to 20% of the explorable range in any direction of U. The regularisation constant C is set equal to 1.

MBDM is applied through the optimisation toolbox of gPROMS ModelBuilder 4.1 employing the solver CVP_SS [20]. The decision functions are identified through the tool for support vector classification implemented in *scikit-learn*, package for machine learning in Python [21].

The experimental design step, illustrated in the proposed methodology in Section 2, will not be considered in the presented case. The design and development of a complete procedure to extended case studies (both in silico and on real setups) is going to be object of future research activities.

4. Results

4.1. MBDM for Parameter Estimation

The available data set was fitted applying both MBDM (14) and a conventional ML estimator (3) for comparing the performance of the two methods. The parameter estimates are given in Table 3 with the associated t-value statistics and the sum of squared residuals, indicated as χ^2. The t-value of reference t_{ref} is also given in the table. This represents a t-value with 95% of significance, obtained from a Student's distribution with degree of freedom equal to the number of fitted measurements. A t-value higher than the t_{ref} is interpreted as an index of satisfactory parameter precision. As one can see from the table, all the computed parameters are statistically satisfactory, but the estimates obtained in the two cases are significantly different. The reason is that in the MBDM case, some of the binary variables β were switched to -1 to satisfy the conditions (13), excluding some experiment from the parameter estimation problem. In Table 4, the binary variables $\hat{\beta}$ computed by MBDM are given for all the

experiments together with the associated investigated conditions. The candidate model was unable to realise low residuals for experiments 4, 8, 12 and 13 (i.e. the experiments with $\hat{\beta} = -1$), which were therefore labelled by MBDM as incompatible with the modelling assumptions. The parity plot in Figure 3a shows the distribution of the residuals achieved by the candidate model if the ML method is employed (i.e. if the whole data set is fitted). In Figure 3b, the residuals associated to the fitted data in the MBDM case (i.e. only the residuals associated to experiments 1-3, 5-7 and 9-11) are reported. The distributions of the normalised residuals associated to the ML method and to the MBDM method are plotted in Figure 4a and Figure 4b respectively. From a comparison of the plots in Figure 3 and the bar charts in Figure 4 one can see that the application of MBDM led to the identification of a model with enhanced fitting capabilities through the automated identification of the experiments causing the bad fitting. The exclusion of experiments 4, 8, 12 and 13 results in a significant reduction of the χ^2, which decreases from 1247.2 in the ML case to 180.3 in the MBDM case.

Table 3: Parameter estimates and related statistics: t-value and sum of squared residuals χ^2; with conventional ML estimator and MBDM estimator.

Method	$[A_1, A_2]$	t-value*	t_{ref}	χ^2
ML	$[5.66 \cdot 10^{12}, 7.33 \cdot 10^{7}]$	$[19.51, 15.39]$	1.66	1247.2
MBDM	$[3.98 \cdot 10^{12}, 6.16 \cdot 10^{7}]$	$[14.63, 11.26]$	1.67	180.3

*a t-value higher than t_{ref} indicates satisfactory parameter precision.

4.2. Domain of expected model reliability

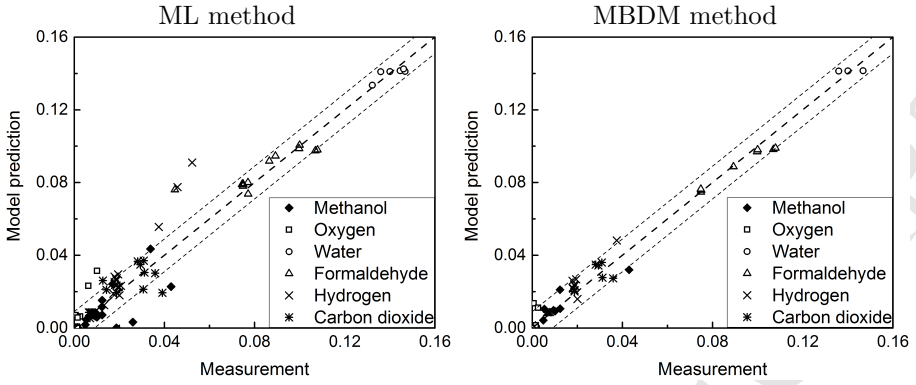
The classification of the experiments as compatible or incompatible with the modelling hypotheses through the binary variable $\hat{\beta}$ is now used to train a SVM algorithm. This leads to the identification of a decision function $I(\mathbf{u})$ in the form of Eq. (17) whose sign provides a classification of the unexplored experimental conditions in terms of good or bad expected model performance. The decision

Table 4: Experimental conditions investigated in the catalytic microreactor and binary variables $\hat{\beta}_j$ computed by MBDM.

N° exp.	T [K]	F* [ml min^{-1}]	$y^{IN}_{CH_3OH}$	$y^{IN}_{O_2}$	$y^{IN}_{H_2O}$	$\hat{\beta}$
1	783	73.1	0.0996	0.0414	0.0754	+1
2	783	41.7	0.0996	0.0414	0.0754	+1
3	783	29.1	0.0996	0.0414	0.0754	+1
4	733	50.9	0.1468	0.0975	0.2293	-1
5	765	50.9	0.1468	0.0975	0.2293	+1
6	796	50.9	0.1468	0.0975	0.2293	+1
7	826	50.9	0.1468	0.0975	0.2293	+1
8	765	93.9	0.1469	0.0980	0.2296	-1
9	796	93.9	0.1469	0.0980	0.2296	+1
10	826	93.9	0.1469	0.0980	0.2296	+1
11	800	54.5	0.2590	0.1064	0.2122	+1
12	850	54.5	0.2590	0.1064	0.2122	-1
13	900	54.5	0.2590	0.1064	0.2122	-1

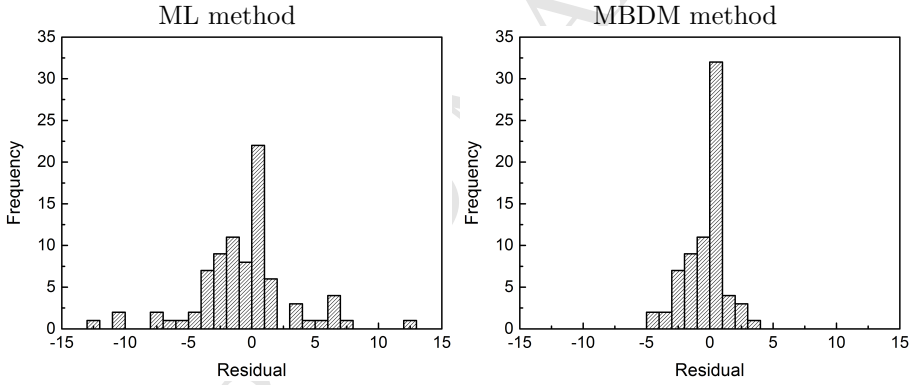
* at temperature $T = 273.15$ K; pressure $P = 101325$ Pa.

function obtained for *Case 1* is represented in Figure 5a in the input subspace defined by temperature and inlet fraction of methanol. Regions of the input space at $I(\mathbf{u}) > 0$ (bright regions in the plot) identify conditions at which the model is expected to provide a good representation of the reacting system. Conversely, conditions at $I(\mathbf{u}) < 0$ (dark regions in the plot) are considered too close to trials that were previously labelled as incompatible. Given a rational choice for the significance α for the $z_{\frac{\alpha}{2}}$ values in (12), in regions at $I(\mathbf{u}) > 0$, the discrepancy between measurements and model predictions is expected to be indistinguishable from measurement noise. The same considerations hold for the decision function identified in *Case 2*, plotted in Figure 5b in the input subspace defined by temperature and inlet fraction of water. Other maps of reliability may be easily computed selecting different sets of training variables,



(a) (b)

Figure 3: Parity plot comparing measurements against model predictions: (a) if a conventional ML estimator is employed; (b) if MBDM is adopted. In (b) only experimental data with $\hat{\beta} = +1$ are reported.

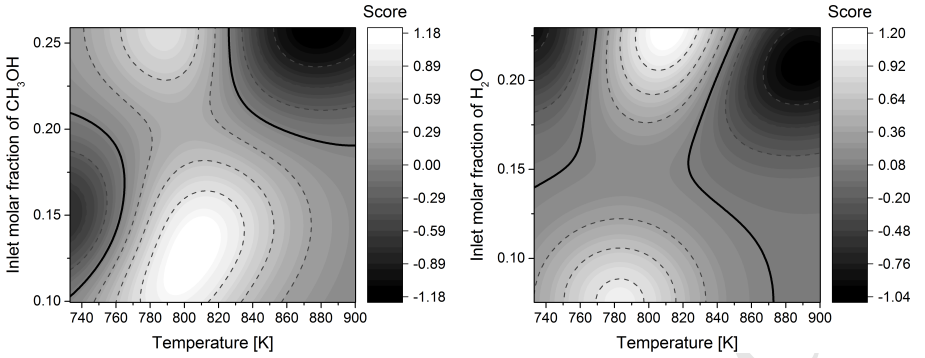


(a) (b)

Figure 4: Distribution of the normalised residuals: (a) if a conventional ML estimator is employed; (b) if MBDM is adopted. In (b) only residuals associated to the experimental data with $\hat{\beta} = +1$ are reported.

possibly involving more than two inputs. Maps of reliability such as those given in Figure 5 may be employed for multiple purposes, e.g.:

- if the approximated model is employed to identify the location of an optimal process point in the input space and the optimum is achieved for



(a) (b)

Figure 5: Score of decision functions identified training the SVM with two different sets of experimental conditions: (a) temperature and methanol fraction at the inlet; (b) temperature and water fraction at the inlet. Solid black lines represent contours at $I(\mathbf{u}) = 0$.

conditions at $I < 0$, one shall proceed carefully and question the reliability of the computed solution;

- if one is willing to enhance the precision on the model parameters collecting new data, the research of optimal experimental conditions through MBDoE methods shall be bounded to regions of the input space at $I > 0$, where the model is expected to provide a good fitting.

Notice that the inclusion of new performed experiments in the data set does not necessarily lead to the computation of different parameter estimates (indeed the new data may be excluded from the parameter estimation by MBDM). However, the inclusion of new experiments in the training set always results in an update of the decision function I and the associated reliability map. In fact, since the decision function in Eq. (17) is influenced by all the experiments available, its score will increase or decrease in the neighbourhood of the experimental conditions associated to the new experiments depending on the labelling computed by MBDM.

The mapping of the design space provided by SVM is purely data driven and depends on the choice of the kernel function as well as the values of the associated

hyperparameters (see Section 2.2). Furthermore, an accurate SVM classification requires the availability of a relatively abundant and distributed training set. Especially at the beginning of the experimental activity, the number of training points may be limited and the classification may be poor. However, notice that in the presented approach for model identification (see Figure 1), an inaccurate classification would only impact the efficiency of the method (i.e. the number of experiments required to identify the model) and not the eventual outcome. An initially inaccurate reliability map may lead to the design of incompatible experiments in regions of the design space that are classified as reliable. However, the accuracy of the SVM classifier increases as the experimental activity proceeds and does not prevent the ultimate identification of an accurate model.

The identified domain of model reliability may be characterised by a very complicated geometry due to the RBF adopted as kernel for the SVM, but also nonlinearity in both the system and the approximated model. The nonconvexity of the region of reliability may result in the achievement of unreliable solutions when employed to bound model-based optimisation problems. In such context, the employment of visualisation techniques may be beneficial for supporting the identification of Pareto optimal solutions [22]. The hybridisation of the proposed model identification approach together with high dimensional data analysis and space visualisation will be object of future studies.

5. Conclusion

The identification of a model with incorrect structure requires both the precise estimation of its parameters and the recognition of the range of experimental conditions where the model can provide satisfactory predictions, namely the domain of model reliability. In this manuscript, a framework for addressing the aforementioned tasks is presented. Fundamental step in the procedure is the fitting of the experimental data through a tailored Model-Based Data Mining (MBDM) method for parameter estimation. MBDM generates two outputs: 1) it computes an instance of the model parameters excluding the data causing the

bad fitting from the parameter estimation problem; 2) it labels the explored experimental conditions in terms of good or bad model performance. The labelled data set generated by MBDM is then used to train a support vector classifier for identifying a decision function to map the space of the experimental conditions in terms of high or low expected model reliability. The identified map of reliability can be employed for raising a flag when optimal process points are identified in the region of low model reliability or to to bound the research of new experimental conditions to investigate for improving the precision of the model parameters. If an optimal process point is identified in a region of low model reliability, the computed solution and the model predictions in its neighbourhood shall not be trusted. At the current stage of the study, the proposed approach does not provide guidance on how to modify the model structure to extend the boundaries of the model reliability domain. It will be object of future work to promote further the integration of machine learning technologies and advanced tools for model building for supporting the development of intelligent algorithms for the quick construction, refinement and statistical validation of phenomenological models.

List of symbols

Latin symbols

a_i	Lagrange multiplier associated to the i-th experiment
A_i	Pre-exponential factor of i-th reaction
b	Hyperplane offset
C	Regularisation constant of the support vector machine
C_i	Concentration of species i
c_{ij}	Arbitrary constant associated to the ij-th element of the log-likelihood
E_{ai}	Activation energy of the i-th reaction
F	Volumetric flowrate
I	Decision function identified by a supervised learning machine
K	Generic kernel function
L	Likelihood function
N_C	Number of chemical species included in the kinetic model
N_{exp}	Number of experiments included in a data set
N_m	Number of dependent output variables in a given model
N_R	Number of reactions involved in the kinetic model
N_u	Number of independent inputs in a given model
N_x	Number of state variables in a given model
N_θ	Number of non-measurable parameters in a given model
P	Pressure
R	Ideal gas constant
r_i	Reaction rate of the i-th reaction
T	Temperature
t_{ref}	t-value of reference computed from a Student's distribution
U	Vector space of model inputs
v	Flow velocity along the axial coordinate of microchannel
y_{ij}	i-th measured variable in the j-th experiment
$\hat{y}_{ij}$	Model prediction of y_{ij}
y_i^{IN}	Molar fraction of species i at the inlet

Z	Feature vector space
z	Axial coordinate of microchannel
$z_{\frac{\alpha}{2}}$	Two-tailed z-value derived from a standard normal distribution

Matrices and vectors

$\mathbf{f}$	Column array of functions $[N_m]$
$\mathbf{u}$	Column array of independent control variables (model inputs) $[N_u]$
$\mathbf{u}_i$	Experimental conditions tested in the i-th experiment $[N_u]$
$\mathbf{x}$	Column array of state variables $[N_x]$
$\hat{\mathbf{y}}$	Column array of predicted output variables $[N_m]$
$\boldsymbol{\theta}$	Column vector of variables representing model parameters $[N_\theta]$
$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$	Maximum likelihood estimate obtained maximising Φ_{ML} $[N_\theta]$
$\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$	Maximum likelihood estimate obtained maximising Φ_{DM} $[N_\theta]$
$\hat{\boldsymbol{\theta}}_{\mathrm{DM,exp}}$	Maximum likelihood estimate obtained maximising $\Phi_{\mathrm{DM,exp}}$ $[N_\theta]$
$\boldsymbol{\Lambda}$	Matrix of binary variables $[N_{exp}\mathrm{x}N_m]$
$\boldsymbol{\xi}$	Array of misclassification errors $[N_{exp}]$
$\boldsymbol{\varphi}$	Vector transformation $\boldsymbol{\varphi} : U \rightarrow Z$ $[\dim(Z)]$

Greek symbols

α	Statistical significance
β_i	Binary variable associated to the i-th experiment in $\Phi_{\mathrm{DM,exp}}$
γ	Decay length of Gaussian radial basis function
Θ	Vector space of model parameters
λ_{ij}	Binary variable associated to the ij-th element of Φ_{DM}
ν_{ij}	Stoichiometric coefficient of the i-th species in the j-th reaction
ξ_i	Misclassification error associated to the i-th experiment
σ_{ij}	Standard deviation of measurement error associated to y_{ij}
Φ_{ML}	Log-likelihood function
Φ'_{ML}	Modified Log-likelihood function
Φ_{DM}	Objective function for model-based data mining of single data

$\Phi_{\mathrm{DM,exp}}$	Objective function for model-based data mining of experiments
χ^2	Sum of squared residuals
Ψ	Data set available for parameter estimation
Ψ'	Reduced data set fitted by a generic data mining method
ψ	Discrete function $\{\mathbf{u}_i,\ i = 1, ..., N_{exp}\} \rightarrow \{-1, +1\}$

References

[1] S. P. Asprey, S. Macchietto, Statistical tools for optimal dynamic model building, Computers & Chemical Engineering 24 (2000).

[2] F. Galvanin, S. Macchietto, F. Bezzo, Model-Based Design of Parallel Experiments, Ind. Eng. Chem. Res. 46 (2007).

[3] Y. Bard, Nonlinear Parameter Estimation, Academic Press, 1974.

[4] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995).

[5] B. Schlkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.

[6] A. J. Smola, B. Schlkopf, A tutorial on support vector regression, Statistics and Computing 14 (2004).

[7] D. Espie, S. Macchietto, The optimal design of dynamic experiments, AIChE J. 35 (1989).

[8] V. Prasad, D. G. Vlachos, Multiscale Model and Informatics-Based Optimal Design of Experiments: Application to the Catalytic Decomposition of Ammonia on Ruthenium, Ind. Eng. Chem. Res. 47 (2008).

[9] J.-L. Dirion, C. Reverte, M. Cabassud, Kinetic parameter estimation from TGA: Optimal design of TGA experiments, Chemical Engineering Research and Design 86 (2008).

[10] F. Galvanin, E. Cao, N. Al-Rifai, A. Gavriilidis, V. Dua, A joint model-based experimental design approach for the identification of kinetic models in continuous flow laboratory reactors, Computers & Chemical Engineering 95 (2016).

[11] J. Bergstra, Y. Bengio, Random Search for Hyper-Parameter Optimization, Journal of Machine Learning Research 13 (2012).

[12] A. Andreasen, H. Lynggaard, C. Stegelmann, P. Stoltze, Simplified kinetic models of methanol oxidation on silver, Applied Catalysis A: General 289 (2005).

[13] F. Galvanin, E. Cao, N. Al-Rifai, V. Dua, A. Gavriilidis, Optimal design of experiments for the identification of kinetic models of methanol oxidation over silver catalyst, Chimica Oggi-Chemistry Today 33 (2015).

[14] J. P. McMullen, K. F. Jensen, Integrated microreactors for reaction automation: new approaches to reaction development, Annu Rev Anal Chem (Palo Alto Calif) 3 (2010).

[15] J. P. McMullen, K. F. Jensen, Rapid Determination of Reaction Kinetics with an Automated Microfluidic System, Org. Process Res. Dev. 15 (2011).

[16] E. Cao, A. Gavriilidis, Oxidative dehydrogenation of methanol in a microstructured reactor, Catalysis Today 110 (2005).

[17] H. Schubert, U. Tegtmeyer, R. Schlgl, On the mechanism of the selective oxidation of methanol over elemental silver, Catal Lett 28 (1994).

[18] M. Quaglio, F. Bezzo, A. Gavriilidis, E. Cao, F. Galvanin, Identification of kinetic models of methanol oxidation on silver in the presence of uncertain catalyst behaviour, AIChE J. (under revision).

[19] E. V. Dokuchits, A. V. Khasin, A. A. Khassin, Mechanism and kinetics of hydrogen oxidation on silver, Russ Chem Bull 61 (2012).

[20] Process Systems Enterprise, gPROMS, www.psenterprise.com/gproms, 1997-2017.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011).

[22] A. Zilinskas, E. S. Fraga, A. Mackute, Data analysis and visualisation for robust multi-criteria process optimisation, Computers & Chemical Engineering 30 (2006).

Appendix A. Conjecture proof

In the present Appendix, a proof is given to demonstrate that optimising the function $\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi)$, i.e. Eq. (A.1), subject to (A.2), is equivalent to optimising the likelihood function $\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi')$, Eq. (A.3), built adopting the reduced data set $\Psi' = \{y_{ij}|\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 1 \,\forall\, i = 1, ..., N_m \wedge\ j = 1, ..., N_{exp}\}$.

$$\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\boldsymbol{\theta})\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \tag{A.1}$$

$$\text{s.t.}\quad \lambda_{ij}(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if} \quad z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2 \geqslant 0 \\ 0 & \text{if} \quad z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2 < 0 \end{cases} \quad \forall \quad i, j \tag{A.2}$$

$$\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi) = \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}})\Big[-\ln(2\pi\sigma^2_{ij}) - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \tag{A.3}$$

The likelihood function $\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi')$ can be rewritten as follows:

$$\begin{aligned}\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi') = {} & \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\boldsymbol{\theta})\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \\ & + \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\big(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\big)\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] \\ & + \frac{1}{2}\sum_{j=1}^{N_{exp}}\sum_{i=1}^{N_m}\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}})\Big[-\ln(2\pi\sigma^2_{ij}) - z^2_{\frac{\alpha}{2}}\Big]\end{aligned} \tag{A.4}$$

The first sum in (A.4) represents $\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi)$ and it attains its maximum in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$. For the second sum, referring for simplicity only to the ij-th element, four cases can be distinguished:

1. the term $\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij} - y_{ij}}{\sigma_{ij}}\Big)^2\Big]$ is non-negative in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ and non-negative in $\boldsymbol{\theta}$. Because of conditions (A.2), $\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 1$ and $\lambda_{ij}(\boldsymbol{\theta}) = 1$, thus $\big(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\big)\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] = 0$;
2. the term $\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij} - y_{ij}}{\sigma_{ij}}\Big)^2\Big]$ is non-negative in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ and negative in $\boldsymbol{\theta}$. From conditions (A.2), $\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 1$ and $\lambda_{ij}(\boldsymbol{\theta}) = 0$, and $\big(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\big)\Big[z^2_{\frac{\alpha}{2}} - \Big(\frac{\hat{y}_{ij}(\boldsymbol{\theta}) - y_{ij}}{\sigma_{ij}}\Big)^2\Big] < 0$;

3. the term $\left[z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}-y_{ij}}{\sigma_{ij}}\right)^2\right]$ is negative in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ and non-negative in $\boldsymbol{\theta}$. From conditions (A.2), $\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 0$ and $\lambda_{ij}(\boldsymbol{\theta}) = 1$, and $\left(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\right)\left[z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta})-y_{ij}}{\sigma_{ij}}\right)^2\right] \leq 0$;
4. the term $\left[z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}-y_{ij}}{\sigma_{ij}}\right)^2\right]$ is negative in $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ and negative in $\boldsymbol{\theta}$. From conditions (A.2), $\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) = 0$ and $\lambda_{ij}(\boldsymbol{\theta}) = 0$, and $\left(\lambda_{ij}(\hat{\boldsymbol{\theta}}_{\mathrm{DM}}) - \lambda_{ij}(\boldsymbol{\theta})\right)\left[z_{\frac{\alpha}{2}}^2 - \left(\frac{\hat{y}_{ij}(\boldsymbol{\theta})-y_{ij}}{\sigma_{ij}}\right)^2\right] = 0$;

Hence, the second sum in (A.4) is always non-positive. The last sum in (A.4) is a constant term and does not depend on $\boldsymbol{\theta}$. It is then concluded that if $\hat{\boldsymbol{\theta}}_{\mathrm{DM}}$ maximises $\Phi_{\mathrm{DM}}(\boldsymbol{\theta}|\Psi)$, it also maximises $\Phi_{\mathrm{ML}}(\boldsymbol{\theta}|\Psi')$.

A model-based data mining approach for determining the domain of validity of approximated models

HIGHLIGHTS

- A framework for the identification of approximated model structures is proposed.
- Machine learning is employed to quantify the model reliability in the input space.
- Reliability maps are given to prevent the use of unreliable optimal process points.
- Reliability maps support optimal design of trials to improve parameter precision.